Correlation Coefficient I

Jozef Hvorecky
Vysoká škola manažmentu / City University
Bratislava, Slovakia

LEVEL
High school or university students with basic knowledge in Mathematics.

OBJECTIVES
To use the calculator's built-in spreadsheet tool to investigate one of basic statistical relationships – correlation between two sets of data – and to demonstrate how it is represented by the correlation coefficient.

Corresponding eActivity
S03CORR.g1e (for Activity 1), S03LEVEL.g1e (for Activity 2)

OVERVIEW
The correlation between two sets of statistical data is explained and discussed. The notions of a strong correlation and of a weak correlation are demonstrated on appropriate data from real-life.

EXPLORATORY ACTIVITIES
[Note]
1. We shall use small letter *x* instead of capital X as shown on the calculator throughout the paper.
2. Some pictures have been artificially processed by a graphic editor to improve their legibility. They then differ from the pictures shown by the calculator display.

Here we describe two activities. For their mathematical background refer for example to [LM], page 343-380.

Activity 1 (S03CORR.g1e):

There is a gold-bearing river near a village. The old mine has been exhausted many years ago, but diggers still occasionally find gold nuggets. Students created a Gold Digger Club. They decided to use their collected gold for their Christmas party. They go out on weekends, pan for gold and keep records on their achievements (see the spreadsheet

table).

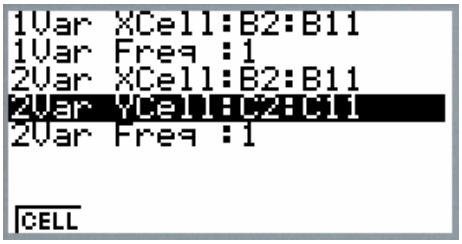
SHEE	A	B	C	D
1	Name	Days	Gold	
2	Ann	14	28	
3	Bob	35	66	
4	David	22	38	
5	Fiona	29	70	
6	Frank	6	22	
7	Kathy	15	27	
8	Mick	17	28	
9	Sarah	20	47	
10	Tim	12	14	
11	Xena	29	68	

As we see from the data, not all club members are equally active. Fiona – the club chair – believes that there is a relationship between the number of visits to the river and the amount of gold collected by the person.
Why is she so confident in the existence of such a relationship? Because she knows the meaning of the correlation coefficient she made its calculation in advance.

(a) (Refer to Correlation) To see whether there is a relationship between two sets of numbers, the data have to be entered into a (spreadsheet) table of the calculator.

In order to perform the requested calculation, press F6 and F2 (CALC). In the CALC submenu, first select F6 (SET) because we have to specify the range of values.

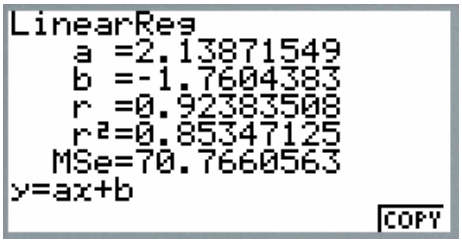
They are B2 to B11 for the x cell range; C2 to C11 for y cell range. Without the proper specification of both ranges, the calculator will either display an error message or use the incorrect data sets and produce false results.



Now the calculation can be executed. Return to the sheet and select F3 (REG).
The next submenu offers a choice of regression methods. From all of them we will only use the simplest one – the linear regression – under F1 (x).

The calculation is executed.

Correlation coefficient I



Pay now your attention to the output value *r* – this is our searched characteristics called the *correlation coefficient*. Its formula is rather complex:

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Again, we have to consider all *n* pairs of elements from two sets x_1, x_2, x_3, …, x_n (the number of visits to gold-bearing river per person) and y_1, y_2, y_3, …, y_n (the number of golden nuggets found by each individual). By combining the numbers in the prescribed method we get the result. Luckily, the calculator performs its evaluation – presumed that we properly specify the ranges of *x* and *y*.

The result is always a real number between -1 and +1 (0.92383508 in this case as shown in the screen dump above). The values -1 and +1 (and values close to them at both ends of the interval) indicate a strong relationship between the compared data sets. The values in the middle of the interval mean weak relationships (with 0 or close to 0 indicating no relationship at all).

(b) (Refer to Fertilizing) During the previous years, a farmer was using a fertilizer in his fields. He kept records on the amount of the used fertilizer (in tons) and yield (in tons of crops). The table shows his records.

SHEE	A	B	C	D
1	Year	Fert	Crop	
2	2000	2	71	
3	2001	1	36	
4	2002	3	79	
5	2003	4	104	
6	2004	2.5	73	

Calculate the correlation coefficient because its result allows us to make a conclusion on reliability of the estimation for the current year (evaluated using the parameters *a* and *b* of the regression line).

Notice that here the exact specification of the range is critical because we have three columns of numbers. Our discussed sets are the amount of the fertilizer (B2 to B6) and the yield (C2 to C6).

EXERCISES A

Exercise 1.
Can we trust to the results estimated by linear regression calculations in the *Fertilizing* problem?

SOLUTIONS to EXERCISES A

Exercise 1.
Yes, we can claim this because the correlation coefficient equals 0.97391263 as seen in eActivity. It is a value very close to 1.

(c) (Refer to Running) On sports event, a medical team measured the time achieved by sportsmen of different age at a "stamina run". All sportsmen started at the same moment and were supposed to jog on a relatively slow speed as long as they could. The medical team recorded the time when each person stopped from exhaustion. The records are in the spreadsheet table.

SHEE	A	B	C	D
1	Age	Time		
2	34	17.5		
3	21	20		
4	27	19.8		
5	18	22.9		
6	42	14.5		
7	33	16		
8	28	20		
9	50	12.2		
10	36	18.3		
11	44	13.2		

Calculate the coefficient of correlation between the two data sets.

EXERCISES A

Exercise 2.
Can we trust to the results estimated by linear regression calculations in the Running problem?

SOLUTIONS to EXERCISES A

Exercise 2.
Yes, we can because the correlation coefficient equals -0.96922327 as seen in eActivity. The value is very close to -1.

Activity 2 (S03LEVEL.g1e):

The strength of relationship can be visualized. Well-correlating data sets form regular patterns clustered around the regression line.

(a) (Refer to Positive correl) Let us return to the gold-panning problem. Display its data using a scattered graph.

EXERCISES B

Exercise 1.
What pattern do the displayed data form?

SOLUTIONS to EXERCISES B

Exercise 1.
The data form a narrow strip from the left lower corner to the right upper corner.

(b) (Refer to Negative correl) Let us return to the stamina-run problem. Display its data using a scattered graph.

EXERCISES B

Exercise 2.
What pattern do the displayed data form?

Exercise 3.
Assume two data sets with a strong correlation. Is there a relationship between their correlation coefficient and their scattered graph?

SOLUTIONS to EXERCISES B

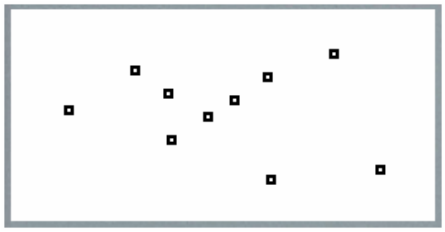
Exercise 2.
The data form a narrow strip from the left upper corner to the right lower corner.

Exercise 3.
The scattered graph of strongly data sets form narrow strips. When the coefficient is close to 1, the strip rises. When it is close to -1, the strip declines.

(e) (Refer to Weak correlation) A group of students collected their data on their height (in meters) and the amount of time they spend on watching TV daily. They are interested whether there is a relationship between the data sets.

SHEE	A	B	C	D
1	Name	Height	Time	
2	Beata	1.42	2.5	
3	Ed	2.02	0.5	
4	Hugo	1.74	2.75	
5	Léa	1.62	1.5	
6	Nick	1.8	3.5	
7	Nora	1.81	0.25	
8	Raj	1.69	2.25	
9	Tanya	1.93	4.25	
10	Tibor	1.55	3.75	
11	Wanda	1.61	3	

The correlation coefficient r = –0.2367005 indicates no important relationship. The same we can observe from their scattered graph.



EXERCISES B

Exercise 4.
Why is the relationship with *r* = -0.2367005 assumed to be a non-significant relationship?

SOLUTIONS to EXERCISES B

Exercise 4.
The relationship with r=0 is rare. A weak relationship varies in a neighborhood of 0. The correlation coefficient of a strong relationship must be very close to -1 or +1.

REFERENCE

[LM] Douglas A. Lind and Robert D. Mason, *Basic Statistics for Business and Economics*, Irwin/McGraw-Hill, 1997. ISBN 0-256-19408-4

